Magyar EuroWordNet projekt: bemutatás és helyzetjelentés

Miháltz Márton

MorphoLogic, Orbánhegyi út 5, 1126 Budapest
mihaltz@morphologic.hu

Kivonat: A tanulmányban bemutatjuk azt a projektet, melynek célja a magyar nyelvű, a EuroWordNet többnyelvű architektúrájába illeszkedő nyelvi ontológia létrehozása. Az ontológia általános része a EuroWordNet-et továbbfejlesztő BalkaNet projekt erőforrásaira épít. Az ontológia kiinduló fogalmi készlete főneveknél és mellékneveknél a BalkaNet Base Concept Set angol nyelvű, Princeton WordNet-ből származó synsetjeinek lefordításával készült, igéknél ezekkel párhuzamosan—a két nyelv igei rendszerének szemantikai különbségei miatt—saját erőforrásokból kiindulva történt. A synsetek lefordítása gépi heurisztikák alkalmazásával, valamint ezek eredményeinek kézi ellenőrzésével történt. A cikkben bemutatjuk az eddigi eredményeket, illetve az ontológia továbbfejlesztésének a projekt során tervezett következő lépéseit.

1 Bevezetés

Természetes nyelvi szövegek gépi feldolgozásában mára vitathatatlanul fontos szerep jutott az ontológiáknak. A legismertebb, nyelvi tudást rendszerező ontológia a WordNet, melyet az 1990-es évektől kezdtek el fejleszteni, először angol nyelvre [9]. A gépi fordítás és az egyéb, több nyelvet kezelő nyelvtechnológiai alkalmazások számára további segítséget jelentenek a többnyelvű, az eredeti angol WordNet anyagához egyéb nyelvű ontológiákat kapcsoló nyelvi erőforrások, melyek első képviselője a EuroWordnet (EWN) projekt volt [14]. Az 1999-ben zárult munka eredménye az angolon kívül 7 európai nyelvre kifejlesztett és összekapcsolt WordNet ontológia volt. A BalkaNet projekt ennek továbbfejlesztése volt 2004-es befejezéssel, további 5 délkelet-európai nyelv bevonásával [13].

Magyar nyelvű, a EWN-hez kapcsolódva többnyelvűséget biztosító, használható méretű és minőségű WordNet ontológia fejlesztésére a GVOP-AKF-2004-3.1.1 pályázati projekt keretei között nyílt lehetőség, három, magyar nyelvtechnológiában vezető intézmény (MorphoLogic Kft., MTA Nyelvtudományi Intézet, Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportja) részvételével, a 2005-2007-ös időszakban. Ebben a tanulmányban ezt a jelenleg is futó projektet szeretnénk bemutatni, az eddigi eredményeket és a további munkát megismertetni.

A továbbiakban először kivonatosan bemutatjuk a WordNet-típusú ontológiák alapfogalmait, a többnyelvű EuroWordNet koncepciót, valamint ennek legutóbbi

megvalósulását, a BalkaNet projektet, melyek kiindulási anyagként szolgáltak a projekthez (2. fejezet). A 3. fejezetben bemutatjuk a magyar EWN ontológia létrehozásának tervezett metodológiáját és a projekt során tervezett lépéseit, majd a 4. fejezetben a cikk elkészültéig megvalósult eredményeit, kitérve az ezen idő alatt felmerült problémákra és a munka közvetlen folytatására.

2 Egy- és többnyelvű wordnetek

2.1 Princeton WordNet

A mentális lexikont, ezen belül az angol nyelv lexikális és fogalmai viszonyait modellező Princeton WordNet (PWN) lexikális szemantikai hálózatot George Miller és munkatársai a Princeton Egyetem Kognitív Tudomány Laboratóriumában, pszicholingvisztikai kísérletek eredményeiből kiindulva fejlesztették ki [9]. A *wordnet* köznév azóta az eredeti, Princeton-ban készült angol nyelvű WordNet felépítését követő nyelvi adatbázisokra utal.

A WordNetben a tartalmas szavak (főnevek, igék, melléknevek, határozószók) különböző értelmeit szójelentéseknek hívják. A szinonimitás jelenségére—egyes szavak bizonyos értelemben, egy adott kontextusban a (denotációs) jelentés megváltoztatása nélkül felcserélhetők—épülnek a synsetek (szinonima-halmazok), a WordNet fogalmi alapegységek. A WordNetben egy fogalom tehát ekvivalens szójelentések halmazával reprezentálható (pl. {léc, deszka}, {fut, szalad, rohan}, {helyes, hibátlan} stb.).

A synsetek között különböző, világismereti, illetve nyelvi kapcsolatot kifejező szemantikai kapcsolatok (relációk) vannak, melyek ezeket a csomópontokat egy összefüggő, irányított körmentes gráfba, fogalmi hálózatba szervezik. A főnévi fogalmak között a legfontosabb reláció a hipernímia (ill. inverze: hiponímia), mely hierarchikus alá-/fölérendeltséget, specifikus/generikus, faj/nem, IS-A öröklődési viszonyt fejez ki (pl. {toll}-{írószer}, {bokor}-{növény}). Speciális altípusa a példánya-hipernímia reláció, mely tulajdonnevekhez kapcsolódó, indivudumoknak megfelelő és általánosabb, osztályoknak megfelelő fogalmak között állhat fent (pl. {Magyarország}-{európai ország}). A hipernímiához hasonló hierarchikus reláció a meronímia (inverzének neve: holonímia), mely rész-egész viszonyt fejez ki. Három fajtája van: egyén-csoport (pl. {fa}-{erdő}), alkotóanyag-tárgy (pl. {cellulóz}-{papír}) és alkatrész-egész (pl. {kerék}-{bicikli}) viszonyt kifejező.

A domain reláció egy tetszőleges fogalom (domain term) és egy témát, fogalmi osztályt (domain) reprezentáló fogalom között áll fenn. Három fajtája van: kategória (szemantikai mező, téma), pl. {teniszütő}-{tenisz}, régió (nyelvhasználók földrajzi helye szerint), pl. {ballup, balls-up}-{ United Kingdom, Great Britain} és használat (nyelvréteg szerinti besorolás), pl. {parázik}-{szleng, argó}. Főnévi fogalmak és más szófajú synsetek között is vannak relációk: tulajdonság mn. és neki megfelelő attribútum fn. között (pl. {piros}-{szín}), morfológiailag rokon (képzett) alakok között (pl. {fekvés}-{fekszik}-{fekvő}).

Főnevek, melléknevek és igék között is értelmezett az antonímia-reláció, mely szembenállást fejez ki valamilyen észszerű denotációs tartományban (pl. {nő}-{férfi}, {megszületik}-{meghal}, {hideg}-{meleg} stb.) Igéknél a hipernímia-

hiponíma relációpárhoz hasonló hierarchikus viszonyt fejez ki a hipernímia-troponímia (pl. {fut, szalad}-{mozog}). Speciális, igei synsetek közötti reláció az előfeltételezést kifejező kapcsolat, pl. {horkol}-{alszik}, illetve az okozás, pl. {meggyújt}-{elég}. Domain, illetve más szófajokhoz kapcsolódó derivációs relációk ezeknél a szófajoknál is vannak. A melléknéveknél a legfontosabb strukturáló reláció az antonímia. A határozószavaknak megfelelő synsetek csak más szófajokhoz kapcsolódnak derivációs morfológiai relációkkal.

A Princeton WordNet jelenleg legfrissebb változata (2.1) mintegy 155.000 angol szót szervez 81.400 különböző főnévi, 13.700 igei, 19.900 melléknévi és 3.700 határozószói synsetbe.

2.2 EuroWordNet

Az 1996-1999 között, az Európai Közösség finanszírozásában megvalósított EuroWordNet (EWN) projekt fő eredménye a WordNet architektúra többnyelvű környezetbe való átültetése volt. A EWN moduláris környezetet biztosít, ahol egy közösen elfogadott közvetítő fogalmi réteghez (Inter-Lingual Idex, ILI) kapcsolódnak a különböző nyelvek (holland, olasz, spanyol és angol, majd német, francia, cseh és észt) wordnetjeinek synsetjei.

Az EWN ILI nagyrészt az angol nyelvű Princeton WordNet 1.5-ös verziójának synsetjeiből állt, a közöttük lévő szemantikai relációk nélkül. Az ún. ekvivalencia-relációk biztosítják az átjárást az ILI-fogalmakon (ún. ILI-rekordokon) keresztül a különböző nyelvek synsetjei között. Ugyanahhoz az ILI-rekordhoz kapcsolt nyelvspecifikus synsetek ekvivalens jelentésűek a nyelvek között. A rugalmas kapcsolat megőrzése érdekében az ekvivalencia-relációnak a pontos azonosság kifejezésén túl több fajtája is van: pl. az adott ILI-fogalomnak egy adott nyelvben csak specifikusabb (vagy általánosabb) megjelenése van stb. Összesen 15-féle ilyen, nyelvek közötti ún. komplex ekvivalencia-relációt definiáltak.

Annak érdekében, hogy a különböző wordnetek fogalmi készlete egységes legyen (általánosságban ugyanazokkal a domain-ekkel vagy fogalmi területekkel foglalkozzanak), a wordneteket egy közösen meghatározott kiinduló fogalmi készlet, a Common Base Concepts (CBC) elemeiből kiindulva építették fel, felülről-lefelé haladva. A CBC fogalmakat a 8 nyelv wordnetjeinek fejlesztői közösen választották ki a PWN synsetjei közül, minden nyelvre lefordították őket, külön-külön kiegészítették egyéb, az adott nyelvben fontosnak ítélt kiinduló fogalommal (Local Base Concepts), és a lokális wordneteket ezekből kiindulva fejlesztették tovább, ahol lehetett, a saját synsteket az ekvivalencia-relációkkal az ILI-fogalmakhoz kapcsolva. A különböző wordnetek tehát egy közös vázra, a Common Base Concept-ekre épülnek, az erre épülő wordnet-struktúrák pedig nyelvenként eltérhetnek.

Noha a teljes ILI strukturálatlan (az angol PWN 1.5 synsetekhez nem vették át a PWN relációit), a részhalmazát képező, CBC 1310 fogalmát egy új, nyelvfüggetlen ontológia, a EuroWordNet Top Ontology (TO) rendszerezi. A TO 63 nyelvfüggetlen fogalom (Top Concept, TC) hierarchiája, melyek fontos szemantikai distinkciókat tükröznek, és meghatározó szemantika-elméletek alapján határozták meg őket. A TO a CBC-ket valójában nem osztályokba szervezi, inkább feature-ök kombinációjaként jellemzi őket, egy CBC-hez több TC is tartozhat. A TC-k a CBC-k ILI-rekordjain keresztül öröklődnek a kapcsolódó nyelvspecifikus jelentésekre.

A EuroWordNet-ben az egyes nyelvek WN-jeit alapvetően két fő különböző módszer egyikével alakították ki:

a) Összevonásos módszer (Merge Model): a lokális alapfogalmakat (BC-k) valamilyen saját erőforrásból kiindulva választották ki, belőlük a synseteket és az azok között lévő relációkat önállóan fejlesztették ki, majd az ekvivalencia-relációkkal leképezték őket az ILI (PWN1.5) synsetekre.

b) Kiterjesztéses módszer (Expand Model): a lokális alapfogalmakat a PWN1.5-ből választották és a PWN1.5 synseteket (kétnyelvű szótárak segítségével) lefordították ekvivalens saját synsetekre. Ebben a megközelítésben a belső relációkat a PWN-ből örökölték, és a továbbiakban, amennyire lehetett, egynyelvű erőforrások segítségével ellenőrizték őket.

Az Összevonásos módszer alkalmazásával a PWN1.5-től független, a nyelvspecifikus tulajdonságokat megőrző wordnetet lehet létrehozni. A Kiterjesztéses módszer a PWN1.5 által erősen determinált wordnetet eredményez. A követendő módszert elsősorban a rendelkezésre álló erőforrások határozták meg.

2.3 BalkaNet

A 2001-2004 között megvalósított, EK finanszírozású BalkaNet (BN) projekt célja a EuroWordNet kiterjesztése volt 5 újabb, délkelet-európai nyelvvel (bolgár, görög, román, szerb és török) [13].

A BN végső változatában az Inter-Lingual Index szerepét a 2.0-ás verziójú Princeton Wordnet synsetjei töltötték be. A BN ILI (BILI) fogalmai fölött egy újabb nyelvfüggetlen ontológiát definiáltak a SUMO felsőszintű ontológia [10] és a PWN fogalmai közötti megfeleltetések átvételével.

A BN-ben a közös kiinduló fogalmi készlet (BalkaNet Concept Set, BCS) 8.516 PWN synsetetből áll: a EWN CBC synsetjein kívül további, az új nyelvek által hozzáadott fogalmakat is tartalmaz.

A projektben az összes erőforrást közös platformra, XML formátumba konvertálták, melyek így a szabadon felhasználható, egyszerre több nyelvi erőforrás böngészését-szerkesztését lehetővé tevő, a BN projekt céljára kifejlesztett VisDic programmal [4] kezelhetők A minőség-ellenőrzőséhez különböző validációs módszereket rendszeresítettek, melyek biztosítják a wordnetek szintaktikai és strukturális helyességét és konzisztenciáját, valamint a nyelvek közötti kapcsolatok érvényességét. [12].

3 A magyar EuroWordNet fejlesztési koncepciója

A bevezetőben említett, 3 éves kutatási-fejlesztési projekt egyik fő terméke a magyar nyelvű EuroWordNet adatbázis lesz. A három intézményből álló konzorcium az ontológia fejlesztéséhez az alábbi stratégiai megfontolásokat fogadta el:

1) A BalkaNet projekt szabadon hozzáférhető erőforrásainak használata.

A magyar wordnet építésének kiindulópontjául nem a EuroWordNet Common Base Concepts, hanem a BalkaNet Concept Set (BCS) synsetjeit választottuk. Utóbbi mellett a következő érvek szóltak:

- A BN BCS a EWN CBC fogalmain felül tartalmaz további 5 európai nyelvben alapvető fontosságúnak tartott fogalmakat (összesen tehát 13 nyelv többnyelvű WN-jének felépítésében hasznosnak tartott információkat, szemben a EWN CBC 8 nyelvével).
- A BCS a Princeton WordNet újabb, 2.0-s verziójára alapul, a EWN CBC a PWN1.5-ösre.
- A BCS 8 516 synsetet tartalmaz, a CBC 1 310-et. A nagyobb mennyiségű synset teljesebb kiindulási alapot ad a létező EWN/BN wordnetek szókincsének magasabb átfedéséhez.
- A BCS fölött rendelkezésre áll két struktúra is (PWN, illetve SUMO hierarchiák), melyek a BN projekt tapasztalatai alapján, rendkívül hasznosak lehetnek az általunk követett kiterjesztéses modell követésekor (ld. lejjebb).

A BCS adaptációjával összhangban az Inter-Lingual Index (ILI) számunkra is a PWN2.0 anyaga lett. Az erőforrásainkat a BN projekt által kialakított XML formátumba konvertáltuk, szerkesztésre és megjelenítésre a VisDic programot választottuk.

2) Ahol lehet, a kiterjesztéses modell, máshol a kiterjesztéses-összevonásos módszerek keverékének alkalmazása.

   Korábbi kísérleteinkben bebizonyosodott, hogy az angol és a magyar főnévi fogalmak rendszere közötti hasonlóság kellő mértékben fennáll ahhoz, hogy a kiterjesztéses módszert követni lehessen [8]. Ennek során a kiindulásul választott ILI (PWN 2.0) synsetjeit automatikus és kézi módszerekkel lefordítjuk és átveszszük a PWN-ben közöttük definiált szemantikai relációkat. Annak érdekében, hogy végeredmény a magyar nyelv fogalmi sajátosságait tükrözze, a lefordított synseteket, illetve az angolból örökölt relációkat alapos kézi munkával, egyenként ellenőrizzük, és ahol szükséges, módosítjuk.

   Amennyiben a nyelvi különbségek miatt ez a módszer tarthatatlan, bizonyos területeken, ill. szófajoknál az összevonásos módszert is alkalmazzuk (magyar synsetek önálló kifejlesztése, beillesztése a magyar ontológiába, majd ILI-rekordhoz kapcsolása).

   Az alapvetően a kiterjesztéses módszert követő megközelítés mellett a megfelelően strukturált erőforrások hiánya, az alacsonyabb fejlesztési költségek, illetve a korábban kifejlesztett és sikerrel alkalmazott, rendelkezésre álló automatikus módszerek szóltak (ld. következő pont).

3) Fél-automatikus módszerek alkalmazása.

   Egy korábbi projekt során olyan módszereket fejlesztettünk ki, melyek segítségével automatikusan lehetett egy magyar-angol alap szótár magyar főnévi címszavait angol (PWN 1.6) synsetekhez hozzárendelni [8]. A 9 különböző heurisztikát alkalmazó algoritmus a kétnyelvű szótárban található strukturális és morfoszemantikai információkon kívül a Magyar Értelmező Kéziszótár [6] egy elektronikus változatából kinyert főnévi definíciókban azonosított szemantikai relációkat használta fel. A különböző heurisztikák eredményeinek legelőnyösebb kombinációja egy kézzel egyértelműsített etalon halmazhoz képest átlagosan kb.

75%-os pontosságot eredményezett (a magyar főnevek és a PWN synsetek között generált kapcsolatokat tekintve).

A legelőnyösebbnek bizonyult automatikus módszereket, újabb erőforrásokkal támogatva (MorphoLogic Tezaurusz szinonimaszótár) alkalmazzuk arra, hogy a BCS, illetve más PWN 2.0 synseteket magyar szinonima-ajánlásokkal lássunk el, melyeket ezután kézi munkával ellenőrzünk.

4) A konzorcium számára rendelkezésre álló szemantikai erőforrások integrációja a készülő ontológiába.

Az automatikusan szinonima-ajánlatokkal ellátott magyar synsetek kézi ellenőrzési fázisa során egyfelől a Magyar Értelmező Kéziszótár (ÉKSz) bejegyzéseit megpróbáljuk megfeleltetni a készülő magyar synsetekkel, másfelől a Nyelvtudományi Intézetben fejlesztett magyar igei vonzatkeret-leíró adatbázis tételeit hozzárendeljük a megfelelő igejelentésekhez.

Az ÉKSz-jelentésekkel létesített leképezés előnye, hogy egyfelől a magyar synsetekhez alkalmas magyar szöveges definíció rendelhető, másfelől az ÉKSz definíciókban feltárt szemantikai relációk alapján lehetőség nyílik az ontológia további kiterjesztésére (ld. 5. fejezet).

5) A BN projekt által kidolgozott minőségbiztosítási metodológia ([12]) adaptációja és rendszeres alkalmazása az eredményeink validációjához. Az ellenőrzésnek a következő kérdéseket kell érintenie:

Synsetek formai ellenőrzése:

- minden synsetben minden literálnak (szinonimának) van jelentés-azonosítója,
- egy synsetben nem lehet két azonos literál (jelentés-azonosítótól függetlenül),
- egy literál ugyanazzal a jelentés-azonosítóval nem fordulhat elő egynél több synsetben (ugyanabban a szófajban),
- egy szó különböző jelentéseihez tartozó jelentés-azonosítók számozása folytonos ,
- literálok automatikus helyesírás-ellenőrzése,
- synset ID ellenőrzés: az azonosítóknak egyedinek kell lennie, minden synset csak a 4 megengedett szófajkód egyikével lehet megjelölve (n, v, a, b)
- synset literáljainak sorrendezése korpuszban megfigyelt gyakoriság alapján.

Belső (lokális) relációk formai ellenőrzése:

- nincsen ugyanaz a reláció ugyanazon 2 synset között többször felvéve,
- a reláció (neve) a standard BN szemantikai relációk (nevének) egyike,
- nincsenek irányított körök,
- a relációkban álló synsetek megfelelő szófajúak,

- egy synsethez kapcsolódó relációknak kompatibiliseknek kell lennie egymással (pl. egy synset nem lehet egyszerre hipernímája és hiponímája ugyanannak a synsetnek),
- nem lehetnek másokkal kapcsolatban nem álló csomópontok,
- minden synsetnek kell, hogy legyen hipernímája, hacsak nem legfelső szintű (gyökér) fogalomnak felel meg,
- nem létező synsetekkel alkotott relációk javítása/törlése.

Synsetekhez tartozó definíciók és használati példák formai ellenőrzése:

- a definíció ne legyen üres,
- a definíció a saját nyelven legyen megfogalmazva,
- definíció szövegének automatikus helyesírás-ellenőrzése,
- a definíció lehetőleg ne tartalmazza a definiált synset literáljait,
- a használati példa a literált a megfelelő szófajjal tartalmazza.

(Az ILI és a magyar synsetek közötti) ekvivalencia-relációk ellenőrzése:

- minden magyar synsetnek legyen(ek) ekvivalense(i) az ILI-ben.

Az ellenőrzések egy része automatikusan elvégezhető, az így felismert hibákat ezután kézzel kell kijavítani.

A magyar WordNet ontológia fejlesztése a következő lépésekben történik: először a kiinduló „mag" részt készítjük el a BN BCS 8.516 synsetjének lefordításával. A fordításhoz a gépi heurisztikákkal minél több angol BCS synsethez automatikus javaslatokat próbálunk tenni. Ezek eredményét emberi munkával, egyenként ellenőrizzük. Az ellenőrzés közben történik az ÉKSz és az Igei Vonzatkeret-adatbázis tételeivel való megfeleltetés, illetve megfelelő magyar synset definíciók és példamondatok megírása. Az összes BCS synset magyar reprezentánsának elkészülte után a PWN-től örökölt szemantikai relációk ellenőrzése-szerkesztése történik egyenként.

A magyarra fordított BCS halmazt ezután kiegészítjük olyan további alapvető fogalmakkal, amelyek nem szerepelnek benne, viszont a kiinduló halmazban fontos szerepük lehet. Ehhez korpuszban (Magyar Nemzeti Szövegtár, illetve ÉKSz definíciós korpusz) megfigyelt gyakorisági értékek alapján keressük meg a potenciális fogalmakat, melyekből synseteket képezünk, azokat beillesztjük a már meglévő ontológiába, valamint megkeressük az ILI ekvivalenseiket. Mindezek után egy teljes validációs ciklussal véglegesítjük a magyar kiinduló fogalmi halmazt (hibák és hibalehetőségek automatikus listázása, kézi javítása).

A második tervezett nagy lépés a mag rész további, felülről-lefelé irányuló kiterjesztése lesz nagy mennyiségű további fogalommal (a végleges magyar wordnet ontológia mintegy 30.000 synsetet fog tartalmazni). A munka részben a BCS létrehozásához hasonlóan fog történni. Automatikus módszerekkel lefordítjuk a Princeton WordNet BCS után fennmaradó synsetjeit, majd az eredményeket a fentiekhez hasonlóan kézzel ellenőrizzük, szerkesztjük és kiegészítjük.

Az angol WN fordítás mellett a már elkészült felső szintek, az ezek és az ÉKSz jelentései közötti megfeleltetések, valamint az ÉKSz definíciókban végzett szemantikai elemezések eredményei alapján lehetőség lesz további magyar hiponímák (troponímák) automatikus hozzáadására [2]. További, fél-automatikus bővítési lehetőséget kínál a nagy mennyiségben rendelkezésre álló tematikus szólisták feldolgozá-

sa (pl. földrajzi nevek, cégnevek, tulajdonnevek stb.) A magyar nyelvben megtalálható derivációs morfológiai relációk a rendelkezésünkre álló morfológiai elemző és generáló eszközök segítségével automatikusan kiegészíthetők. Mindezen munkák eredményét szintén kézzel kell majd ellenőrizni.

Az utolsó lépés egy speciális, üzleti szakkifejezéseket tartalmazó szakontológia kifejlesztése és az általános ontológiához kapcsolása lesz. Az üzleti szakontológia a projekt többi célkitűzéseit (információ-kinyerés többmondatos rövid szöveges üzleti hírekből) fogja támogatni [1]. Elkészítésében támaszkodni fogunk a szabadon hozzáférhető Teknowledge Financial Ontology szakontológiára[15], valamint a MorphoLogic rendelkezésére álló angol-magyar üzleti szótárakra.

4 Jelenlegi eredmények

A munka kezdetén a korábbi hasonló projektek ([2], [3]) eredményei alapján kifejlesztett, automatikus synset-fordító heurisztikákat ([8]) alkalmaztuk a 8.516 db BCS synset magyarra fordításához. Az alábbiakban röviden bemutatjuk a kiválasztott heurisztikák algoritmusait és a támogató nyelvi erőforrásokat:

a) Egyjelentésű angol szavak: ha egy magyar szó valamelyik angol fordítása egyértelmű a WN-ben, vagyis csupán egyetlen synsetbe tartozik, akkor létrehozunk egy kapcsolatot a magyar szó és a synset között.

b) Többjelentésű angol szavak egyértelmű fordítással: ha egy angol szónak csak egyetlen, egyértelmű magyar fordítása van (a magyar szónak csak ez az egyetlen angol fordítása), és az angol szó a WN-ben több synsethez is tartozik, a magyar fordítást hozzárendeljük ezekhez.

c) Variánsok: ha egy WN synset kettő vagy több olyan angol szót tartalmaz, melyeknek csupán egyetlen magyar fordításuk van, és az ugyanaz a magyar szó, akkor a magyar szót hozzárendeljük a közös synsethez.

d) Szinonimák: a magyar szó angol fordításaihoz tartozó synsetek közül azt választjuk ki, amely a legtöbbet tartalmazza a szó szinonimáinak angol fordításai közül (de legalább kettőt). Magyar szinonimák előállításához felhasználtuk egyfelől az ÉKSz definíciókban géppel azonosított szinonimákat ([8]), másfelől a ML Tezaurusz szinonimáit.

e) Latin nevek: ha egy magyar szóhoz rendelkezésre áll latin megfelelő (állat- és növényfajok, rendszertani kategóriák stb.), akkor azt az angol synsetet választjuk, ami az angol fordításon kívül a latin nevet is tartalmazza. Latin ekvivalenseket az ÉKSz-ből, illetve a kétnyelvű szótárakból azonosítottunk magyar címszavakhoz.

f) Minimális fogalmi távolság: amennyiben van egy magyar szó és egy hozzá tartozó magyar hiperníma szó, akkor képezzük ezek fordításainak lehetséges

[15] http://ontology.teknowledge.com/

synsetjeit, majd belőlük megkeressük azt a párt, ami a WN fogalmi hálózatában a legközelebb helyezkedik el egymáshoz. A magyar címszót a minimális távolságú pár megfelelő tagjához rendeljük.

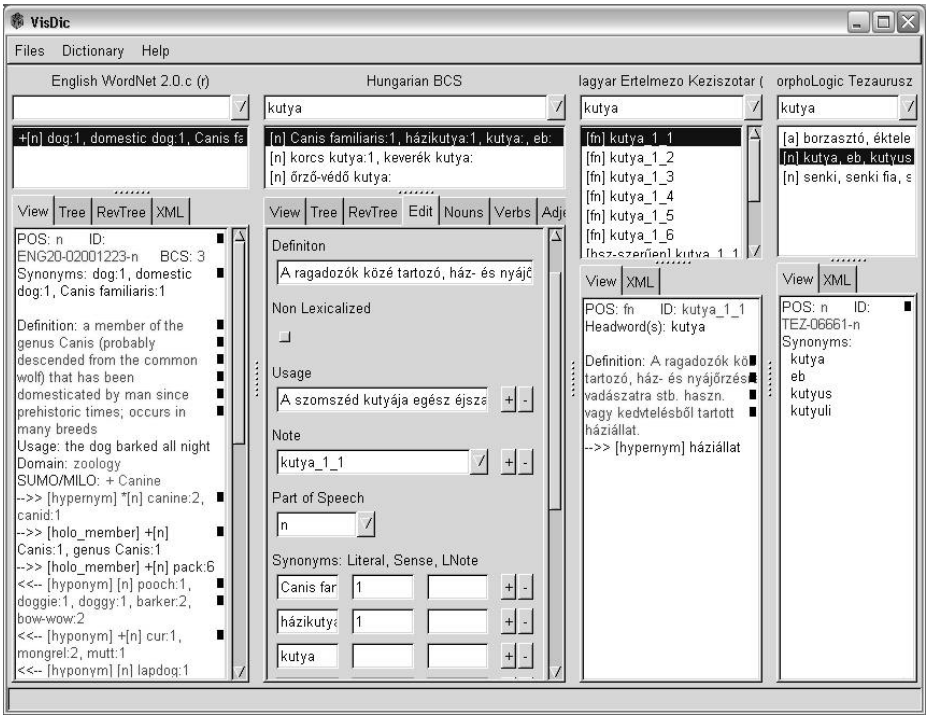
Magyar címszavakhoz hiperníma szavak kétféle forrásból álltak rendelkezésre: egyrészt az ÉKSz definíciók szemantikai elemzésével, másrészt a kétnyelvű szótárak összetett főneveinek morfológiai elemzésével. Ez utóbbi kétféleképpen történhetett: endocentrikus főnévi-főnévi összetételek (egybe írt összetett szavak) utótagjának azonosításával, illetve többszavas lexikalizált főnévi kifejezések esetén a fej azonosításával.

Mivel a BCS synsetek csak kb. 87%-ában volt legalább egy szinonimának magyar fordítása, az automatikus fordítás által elérhető elméleti maximum lefedettség 87% volt. A heurisztikák kombinált eredményei a teljes BCS anyagának kb. felét voltak képesek lefedni (1. Táblázat).

1. Táblázat: a BCS automatikus fordításának eredményei

	BCS	**Automatikusan. lefordítva**	**(%)**
Főnévi synsetek	5 896	3 149	(53,41%)
Igei synsetek	2 318	1 139	(49,14%)
Mn-i synsetek	302	77	(25,50%)
összesen:	*8 516*	*4 365*	*(51,26%)*

Az automatikus fordítást a kézi ellenőrzési-szerkesztési fázis követte: a lefordított synsetek szinonimáit ellenőrizni és/vagy kiegészíteni, a nem lefordított synseteket pedig le kellett fordítani. A munkához segítséget a VisDic-be integrált Értelmező Kéziszótár és a MorphoLogic Tezaurusz nyújtott. A szinonimák ellenőrzése mellett az ÉKSz, illetve a NYTI vonzatkeret-adatbázissal való összekapcsolás is ekkor történt. A munka egy 2 oldalas irányelv betartásával, a VisDic editor felhasználásával folyik (1. Ábra). A cikk megírásának időpontjáig a BCS mintegy háromnegyedét dolgoztuk fel.



1. Ábra: 4 szótár szimultán használata a VisDic program segítségével. Balról jobbra: Princeton WordNet 2.0, synset áttekintő nézet; magyar WordNet, synset szerkesztés nézet; Értelmező Kéziszótár és ML Tezaurusz, szócikk megjelenítés. Az automatikus szinkronizálásnak köszönhetően az angol és a magyar wordnetek ekvivalens fogalmakat jelenítenek meg. A másik két szótárban kézzel kerestük meg a megfelelő szócikkeket.

A manuális munka során számos, előre nem látott nehézséggel és problémával találkoztunk.

Először is, az igei rész fordításának elején nyilvánvalóvá vált, hogy a kiterjesztéses módszer nem tartható teljes mértékben. Egyrészt az eredeti PWN-ben az igék rendszerében megfigyelt hibák és inkonzisztenciák, másrészt a két nyelv igei rendszere közötti morfológiai és szemantikai különbségek felismerése miatt úgy döntöttünk, hogy az igei részt nem teljes mértékben az angol taxonómiára támaszkodva, hanem részben önállóan elindulva készítjük el (részletesen ld. [7]).

A deverbális (igékből képzett) főneveket tartalmazó synsetek feldolgozásakor hasonló okokból szintén problémákat észleltünk. Pl. ezen a fogalmi területen kiugróan magas azon BCS synseteknek a száma, melyek nem, vagy csak igen nehezen, csak nem lexikalizált, ad hoc frázisokkal fordíthatók magyarra. Mindezek miatt célszerűnek látszik a közeljövőben a deverbális főnévi és az igei rész taxonómiáját összehangolni (a deverbális synsetek közötti relációkat a morfológiailag megfelelő igei synsetek között kézzel megállapított relációkhoz igazítani). A derivációs relációkat automatikusan kell majd hozzáadni.

A biológiai taxonómikus fogalmak (állatfajok, -nemek, -rendek stb.) fordítása során előforduló problémák közül kiemelhető, hogy sokszor a WN synsetek pontatlanok voltak, illetve a segítségként felhasznált magyar taxonomikus forrásoktól eltérő

rendszerezéseket mutattak be. Ezen synsetek lektorálásához szeretnénk egy biológus szakfordítót felkérni.

Sokszor előfordult, hogy egy PWN synset lefordításakor nehezen lehetett pontosan azonosítani a fogalmat, mivel a synset hipernímája (hiponímája) nagyon apró, nem anyanyelvi beszélő számára csupán nagyon nehezen észrevehető különbséget tartalmazott a kérdéses synsethez képest. Az ilyen eseteket megjelöltük a relációk kézi ellenőrzési munkaszakaszához. A megoldást a komplex ekvivalencia-relációk alkalmazása jelentheti, melyekkel modellezhetők azok az esetek, amikor a magyar hierarchia nem követi pontosan az angolt.

Részletes, összefoglaló adatokat és felvetéseket a teljes BCS anyag fordításának elkészülése és elemzése után tudunk megadni, amit egy későbbi publikációban tervezünk bemutatni.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas Gy.: Construction of the Hungarian EuroWordNet Ontology And Its Application To Information Extraction. To appear in: Proceedings of the 3rd International WordNet Conference, Jeju Island, Korea (2006)
2. Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997)
3. Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)
4. Horak, A., P. Smrz: New Features of Wordnet Editor VisDic. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
6. Juhász, J., I., Szőke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest (1972)
7. Kuti, J., Vajda P., Varasdi K.: Javaslat a magyar igei WordNet kialakítására. Elbírálás alatt a III. Magyar Számítógépe Konferencián, Szeged (2005)
8. Miháltz, M.: Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, January 20--23 (2004)
9. Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.
10. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19 (2001)
11. Prószéky, G.: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany, pp 149–158 (1996)
12. Smrz, P.: Quality Control and Checking for Wordnets Development: A Case Study of BalkaNet. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
13. Tufiş, D., D. Cristea, S. Stamou: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)
14. Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document (Deliverable D032D033/2D014) (1999)